Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference

Abstract

In recent years, applying multi-modal large language models (MLLMs) in various fields has achieved remarkable success. However, as the foundation model for many downstream tasks, MLLMs comprise the well-known Transformer network, which has a less efficient quadratic computation complexity. In this study, we introduce Cobra, a multi-modal large-scale language model built upon a state-space model, which has demonstrated significant potential in efficiently handling long sequences with fast inference and linear scalability concerning sequence length. Specifically, Cobra involves replacing Transformer-based backbone models (e.g., LLaMA or Phi) with pre-trained Mamba language models. We then empirically explore effective strategies for aligning visual and textual modalities and integrating various pre-trained Mamba model variants with visual encoders. Experiments across various multi-modal benchmarks demonstrate that: (i) Cobra performs 3x ~ 4x faster than the most computationally efficient state-of-the-art methods, e.g., LLaVA-Phi and MobileVLM v2. Additionally, its performance is significantly enhanced thanks to the implementation of linear sequential modeling. (ii) Cobra fine-tunes a small parameter (~48% of model parameters), leading to a significant improvement in overall performance compared to LLaVA.